图书情报工作
第66卷 第20期 2022年10月

大数据环境下知识融合技术体系研究*

■ 陈沫　李广建
北京大学信息管理系　北京 100081

摘　要：[目的/意义]研究在大数据环境下多源知识融合的关键技术，结合不同领域多源知识对象的特点构建出一整套技术体系，为知识融合的落地实现提供技术支撑和解决方案。[方法/过程]运用定性分析法对现有的相关研究进行分析，再对相关内容进行归纳和演绎，利用文献分析法，梳理知识融合要解决的问题，归纳总结知识融合的任务类型以及实现各种任务所需的工作流程及其涉及的具体技术，形成知识融合技术体系。[结果/结论]综合考虑各种技术的自身特性、适用的知识对象、应用的抽象层次，建立具有计算层、功能层和任务层三个层次的技术体系架构。这三个层次相互联系、相互影响、环环相扣，向上可以抽象，可以与知识融合的具体问题（任务）相关联；向下可以具化，即找到解决知识融合具体问题的可操作、可计算的技术方法。

关键词：知识融合　技术体系　大数据
分类号：G250
DOI：10.13266/j.issn.0252－3116.2022.20.003

1　引言

科学技术的迅猛发展大大增加了人类的知识总量，大量的知识分布在世界上不同的数据源中，并以不同的形式呈现。在信息化社会，知识已经成为引领经济增长的战略要素，知识积累能够驱动经济的增长，而相关技术的领先程度和创新能力直接影响到知识优势的形成[1]。因此，如何在技术层面上实现知识对象的融合已经成为大数据环境下知识融合的基础和需要优先解决的问题，这涉及从资源的收集存储、计算分析到访问应用等多个层面的技术和方法。

本文所说的大数据环境下的知识融合技术，包括两层含义：

（1）它是为解决知识融合问题而形成的技术。具体地说，它是指解决各类知识融合问题的技术簇，并不是一类单一技术，而是具有递归结构的、模块化的技术组合，通过对各类基本技术组件或技术单元的组配，能够最终实现面向特定任务的、特定功能的、多尺度的知识融合。大数据时代，知识融合需要的技术的特征与应用流程有其自身的规律，反映了知识融合这一特殊现象的内在要求。同时，随着对知识融合研究的不断深入以及对知识融合要求的不断提高，各类信息技术也会针对知识融合自身的特征而有所改进和提升，甚至需要专门针对知识融合研发出特定的实现技术。这里的知识融合技术包括了可以应用（直接应用和适应性改造后应用）于知识融合过程实现的信息技术以及专门为知识融合而研发出的实现技术。

（2）它本身具有大数据技术的特点。大数据环境下，知识融合面对的数据体量庞大、密度低、实时性强、结构多元、来源多样，传统的知识融合技术与方法难以处理，需要依托于新型计算架构。针对数据体量庞大，大数据环境下的知识融合技术要支持对超大数据集的拆分，将它们分散到多台机器上，利用多台机器的存储和计算能力共同完成任务；针对数据密度低、实时性强，大数据环境下的知识融合技术从单纯的"数据集"处理转向了具备"数据流"处理的能力，也即支持对大规模流动数据（流数据）在不断变化的运动过程中的实时处理；针对数据结构多元、来源多样，大数据环境下的知识融合技术能够处理多种类型的数据，包括数值、文本、图形图像、音视频等等。更具体地说，目前常说的大数据技术，基本上都是属于 Hadoop 生态体系中的技术，因此，也可以将大数据环境下的知识融合技术理解为 Hadoop 生态体系中可以解决知识融合问题的那些技术。它们的来源主要有两个，一个是对传统知

* 本文系国家社会科学基金重大项目"大数据时代知识融合的体系架构、实现模式及实证研究"（项目编号：15ZDB129）研究成果之一。
作者简介：陈沫，博士研究生；李广建，教授，博士，博士生导师，通信作者，E-mail：ligj@pku.edu.cn。
收稿日期：2022－07－29　修回日期：2022－08－26　本文起止页码：20－31　本文责任编辑：易飞

陈沫，李广建. 大数据环境下知识融合技术体系研究[J]. 图书情报工作,2022,66(20):20－31.

识融合技术进行升级改造而形成的符合 Hadoop 生态规范的知识融合技术,另一个是在 Hadoop 生态体系中生长出的知识融合技术。

现有有关知识融合技术方法的研究,无论是对于传统的知识融合技术方法还是对大数据环境下的知识融合技术方法的研究,主要是为知识融合活动提供了底层的技术支撑,但普遍较为零散,对个别操作层面的针对性较强,即便是有涉及知识融合技术体系的研究,也基本上是对关键技术的简单分类,缺乏对知识融合技术体系的整体研究,如 X. Yu 和 Q. Lin[2] 将知识融合方法分为 4 种:基于本体、基于规则、基于统计学习和基于上下文语境的知识融合方法;与之相似,邱均平和余厚强[3] 根据实现路径将知识融合技术分为四大类:基于语义规则、基于贝叶斯网络、基于 D-S 理论和基于知识挖掘的知识融合;缑锦[4] 总结得出了基于遗传算法、基于融合规则、基于改进算法和基于需求驱动型算法的知识融合方法分类体系。

大数据环境下,知识融合更加强调多种来源的知识融合、多种结构和表现形式的知识之间的融合。即使不考虑大数据生态规范方面的特征,知识融合技术的应用也面临着新的要求,需要综合利用多种现有的知识融合技术,并将这些底层技术方法与具体融合实践相结合。根据知识融合需求和融合任务的不同,或者在某种底层技术方法的基础上进行参数调整,或者将多种底层技术方法加以"组合"和"集成",使它们相互配合解决融合问题,甚至在知识融合实践中创造出全新技术方法。为此,需要我们对大数据环境下知识融合的技术体系得出完整的认识,使其能够容纳各种知识融合技术(包括容纳目前未知的知识融合技术)并明确各种技术对知识融合可能起到的作用,以适应大数据环境下知识融合的新特点。

本文拟结合不同领域多源知识对象的特点,构建大数据时代知识融合技术体系,为知识融合的落地实现提供技术支撑和解决方案,也为知识融合的技术理论和技术应用提供参考。为清楚和简化起见,本文在下文中大部分情况下省略了知识融合技术之前的"大数据环境下"。

2 知识融合技术体系构建

知识融合技术发端于专家系统,成长于互联网时代,成熟于大数据时代。大数据时代数据与知识的爆炸性增长,使得各方对知识融合的应用需求也愈发强烈,知识融合技术也随之蓬勃发展,信息科学、医学、工程等多个领域也开始重视本领域中知识融合技术的研究。由于知识融合本身是一个应用性很强的研究领域,以往知识融合技术研究往往从某一学科或某一领域的角度出发,以解决本学科或本领域中涉及的知识融合的具体问题为着眼点展开研究,到目前为止尚未形成一套成熟的、统一的知识融合技术体系。这里,首先简单介绍本文构建知识融合技术体系的基本思路,而后提出知识融合技术体系框架。

2.1 知识融合技术体系及其构建思路

体系泛指一定范围内或同类的事物按照一定的秩序和内部联系组合而成的整体,是由不同子系统组成的系统。技术体系则是技术整体性的表现形式,是一种宏观的、社会性的整体技术结构[5],一般是由各种技术有机联系而形成的具有特定功能的统一体。大数据时代的知识融合技术体系,是指能够反映大数据环境下知识融合所使用的各类信息技术及其相互关联关系的统一整体。

如前所述,知识融合技术不是指一类或一种单一的技术,而是一个技术簇,从字面上理解,它是由基于技术内在关联性和差异性而形成的技术集合或技术体系。每项具体的技术在这个技术集合或技术体系中的地位和作用是不一样的,这也就导致了从不同角度对技术这一术语的不同理解,至少有以下几种:技术是工具或手段;技术是方法或者是关于方法的知识;技术是人类活动(过程)或人类行为;技术是技能、方法、手段、工具和知识的某种组合或总和[6]。由此可见,当人们在使用知识融合技术这一术语时,往往有着不同的层次,既可能指的是完成某项知识融合任务的路径,即根据知识融合任务的要求,将已知的手段、方法或工具组合起来,甚至根据要解决的问题发明创造新的手段、方法或工具,形成解决问题的总体框架;也可能指的是对所要融合的对象(数据和知识)进行处理的操作过程,包括操作步骤、规则和技巧。由此可见,构建知识融合技术体系的过程,就是要梳理不同层次的知识融合概念,明确它们在知识融合技术簇中的地位及相互关系,反映它们在解决知识融合问题过程中相互配合、相互补充的特性。

借鉴前人对技术这一概念的理解,结合本文构建知识融合技术体系需要,本文将知识融合技术归纳为三个层次:计算层次、功能层次以及任务层次。计算层次技术是最基础的操作技术,它们面向具体的数据,涵盖了知识融合过程处理数据的公式、变量、模型、指标、算法等等,体现的是"技术是手段、技能和方法"这一

含义。功能层强调技术在知识融合过程中发挥的作用,它是指为完成知识融合任务所要解决的单元问题,任何一项知识融合任务都可以分解成需要具体解决的问题,这些问题的组合构成了知识融合的任务。功能层的技术起着承上启下的作用,一方面,它接收来自计算层对数据处理的结果(也即综合了计算层的多种技术);另一方面,功能层内的不同技术的组合,对计算层数据处理的结果进行了融合,解决了具体知识融合功能,因此,功能层次的技术大致体现的是“技术是关于方法的知识”或是“人类活动(过程)或人类行为”这样的含义。任务层的技术主要反映了知识融合主体在完成知识融合任务中的主体性,是功能层技术的组合或叠加,它体现的是技术是“方法、手段、工具和知识的某种组合或总和”。这样构建起来的知识融合技术体系,既可以反映出知识融合技术的可操作性特征,也可以反映出知识融合技术在解决知识融合任务时的主观能动性,即通过对已知可操作技术的组合,创新性地实现知识融合任务。而对于任何一项知识融合任务,只要明确总体解决问题的思路和框架,技术的使用者都能找到与完成任务相关联的公式、变量、模型、指标、算法,从而完成知识融合任务。

为了充分揭示知识融合技术体系中不同层次技术的多方面特征,避免仅仅从单一的(标准)来区分和类聚技术,防止发生遗漏造成列举不全,本文使用多维分类方法构建大数据环境下的知识融合体系。多维分类是从多角度属性对事物或对象进行分类,用多种特征或组面的系列来细分类目,它可以在对事物或对象作相应属性解析的基础上,采用若干个可并列的属性维概念来对事物或对象进行多重立类和类别划分,并可在计算机技术支持下构建多类别组配的网状结构,从而达到多角度地认识事物或对象的目的,并且可以随着事物或对象的增减而调整属性维度(分面),不会影响原有的体系结构。在每一种技术层次中,针对该层次的技术特征可以设定不同的维度(分类标准)表,维度表下还可依据相关性划分为多个层级,层级是对维度的进一步细化,多维度、多层级的划分才能深入分析与表达知识融合技术的具体内容与特征。不同维度及不同层级的方法都是可以多重出现的,这种多维分类体系可以对知识融合方法相关内容属性及外部特征进行多角度网状式揭示。

2.2 知识融合技术体系的整体框架

根据以上构建思路,本文从知识融合的任务层、功能层和计算层 3 个层面构建大数据环境下的知识融合技术体系,如图 1 所示:

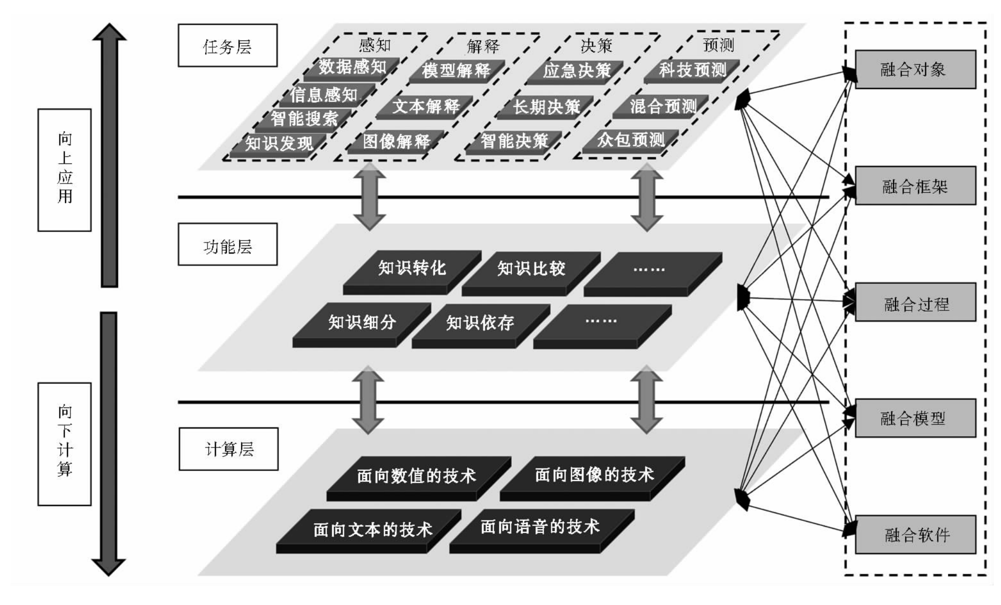


图 1 大数据时代知识融合技术体系框架

2.2.1 计算层技术

计算层是知识融合技术体系中最为具体的底层,它主要提供各种技术实现层次的内容,例如算法流程、实施模型、实现公式等,这个层次中的每一种具体的技术,通常只是实现或者是完成知识融合过程中的某一个环节的一部分工作,因此,计算层知识融合技术是实现知识融合的技术基础,无论是多么复杂的知识融合任务的完成以及知识融合功能的实现,最终都要落实到这些可实现的算法流程、实施模型和实现公式上,任何复杂的知识融合功能和复杂知识融合任务的达成,

都需要一种或多种(主要是多种)计算层知识融合技术的有机组合和相互配合。计算层知识融合技术还可以利用不同的标准做进一步的细分。

2.2.2 功能层技术

功能层是指在知识融合实践过程中,为了完成知识融合任务,用于知识融合过程中各个功能环节中的各种技术。计算层中具体技术的组合叠加实现了知识融合的具体功能。关于知识融合的功能始终都是知识融合领域研究中的重要内容,无论是对知识融合概念还是知识融合系统的研究,都离不开对知识融合功能的认识。知识融合可以被视为一个以知识为操作对象的复杂系统,而系统所具有的功能是指其在不同的方面所能够提供的功用、功效或用途,所以可以通过功能分解的方式来描述该系统。徐敏认为,人类处理知识无非就是解决知识形式的抽象和转换、知识的化繁为简、发现知识内容的相同与差异以及找出知识中的联系与关系等问题,据此将知识融合的功能概括为知识转化、知识细分、知识比较和知识依存[7],本文沿用这种知识融合功能的分类并对完成各个功能所需要的各种具体实现技术进行了论述和总结,明确了功能层技术与计算层技术的对应关系。

2.2.3 任务层技术

任务层的目标是从解决具体任务的角度去审视知识融合技术,这里所说的任务是指知识融合为解决具体问题而进行的各种有目的的活动。任务层中的技术是功能层和算法层的有机组合,多个功能共同组合完成一个任务,是技术体系的最高层。随着知识融合研究对象的不断扩展,知识融合产生了多个与学科或领域高度相关的应用分支,不同学科或领域的知识融合的具体任务也不同。由于各个具体的学科或领域有自身的特点,不可能一一列举它们所有的知识融合任务,但是,这些学科或领域的知识融合是有共性的,也就是说,抛开具体学科或领域中的知识内容,在解决学科或领域问题时所开展的活动是类似的,所要用到的技术也是基本共通的,如果说有特色的话,那就是要根据本学科或领域中的知识内容对技术做适应性改造。本文中,结合前人的研究[8－11],将知识融合的任务粗粒度地划分为感知、解释、决策和预测。由于知识融合中每一项任务所要解决的问题不同,所需要的核心技术或主要技术也会有所不同。

在上述知识融合技术体系中,计算层、功能层和任务层3个层次中的技术均不是独立存在的,三者形成了一种递进关系,是相互联系、相互影响、环环相扣、缺一不可的。在实际应用中,可以从最顶层出发,从任务层选择抽象的任务,进而具象化到中间层的功能选择,最后向下选择计算层的具体算法,从而为实际的知识融合场景进行技术支撑。反之,也可以从计算层中的具体技术入手,逐层向上抽象,了解具体技术可以参与实现的功能以及参与实现的任务。

下文中,对知识融合技术体系中的计算层技术、功能层技术和任务层技术中的内容加以详细说明。

3 计算层知识融合技术

在计算层技术中,强调技术的可操作性,具体地说,这一层次中技术这一术语是指可以具体实现的算法、公式、模型或指标。这些算法、公式、模型或指标的处理对象是各种类型的具体数据,对不同数据的处理需采用不同的技术来满足不同的数据特征,因此,将处理对象的类型作为计算层知识融合技术的分类标准,是最适合的计算层技术划分方式。由此,将计算层知识融合技术分为面向数值型数据的技术、面向文本型数据的技术、面向图像型数据的技术与面向语音型数据的技术等,如图2所示:

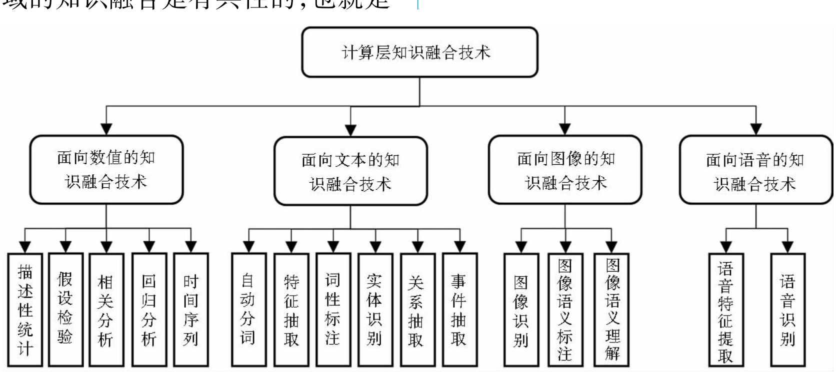


图2 计算层知识融合技术体系

3.1 面向数值型数据的知识融合技术

知识融合的早期处理对象就是数值型数据,最早的数据融合就是面向诸如传感器数据之类的数值型数据的融合,许多面向其他数据类型的知识融合最终也需将数据转化为数值型数据后再进一步做融合和分析。因此,数值型数据是知识融合过程中不可或缺的重要数据类型。面向数值型数据的知识融合技术是从大量的、异构的、不完全的、有噪声的、模糊的、随机的数值型数据集中识别出有效的、新颖的、潜在有用的数据,进行一系列操作以达到融合的目的。一般而言,面向数值型数据的知识融合技术可以大致可以分为两类,一类是统计分析的技术,一类是挖掘预测的技术。

统计分析技术刻画目标数据中数据的一般性质,对某一主体或现象的数据或信息进行整理、计算、构建指标体系,对研究目标进行整体的概括理解,找出规律。统计分析技术是知识融合的一项基础技术,绝大部分知识融合功能的实现都离不开统计。知识融合中常用的统计分析技术有描述性统计、假设检验、相关分析、回归分析和时间序列分析等,如贝叶斯方法(Bayes)[12-13]、卡尔曼滤波算法[14],正如 G. Feng 等[15]提到的那样,综合贝叶斯网络是实现知识融合非常有潜力的技术。

挖掘预测技术从大量数据或信息中提取、揭示或归纳出有价值的潜在内容,构建及验证事物关系、类别及关联性,并找出隐含在数据中的规律性。知识融合中常用的技术有联机分析、分类、聚类和关联规则等,如蚁群挖掘分类技术[16]、基于扩展概念格的多数据源分类知识融合[17]、基于监督学习的多源知识融合方法[18]。

3.2 面向文本型数据的知识融合技术

面向文本型数据的知识融合是目前各领域中最常见的知识融合类型。文本型数据多为非结构化的数据,在文本型数据的知识融合中,通常会以篇章、段落、短语甚至文章中的实体等为单位进行知识的抽取、挖掘和推理。例如,在热点识别的研究中,关键词的共词分析就是最经典的方法之一,该方法通过建立共词矩阵和设置一定的阈值,能够将科技报告、专利文献和学术论文等多种来源的文档进行融合,发现文档集中最为高频出现的一系列关键词语,再将它们进行归纳整合,从而发现科学技术研究的热点或前沿。

面向文本型数据的知识融合技术主要用来对大规模文本库进行挖掘分析,完成文本主题识别、知识体系构建、知识异同的比较、知识之间的关联等。面向文本型数据的知识融合技术主要包括自动分词、文本特征抽取、词性标注、命名实体识别、关系抽取、事件抽取、主题建模与文本表示、文本挖掘与模式发现等,如利用关联规则抽取[19]、文本分类[20]、基于语义关系和规则来融合分布式知识对象的知识融合算法[21]等。

3.3 面向图像型数据的知识融合技术

计算层中面向图像型数据的知识融合技术涉及了图像处理技术的多个部分,包括图像增强和复原、图像分割、图像识别、图像描述、图像压缩和解压等。图像数据的知识融合,主要有两种应用场景,一种应用场景是多个图像之间的融合,通常是将来自同一对象不同时间的图像或者来自两个或两个以上的图像采集设备在同一时间(或不同时间)所获得的同一对象的图像进行综合,消除多个图像中可能存在的矛盾、冗余、不一致等现象,从而获得关于图像的新的知识[22-23];另一种应用场景是与文本、语音等数据进行融合,从而加深对研究对象的理解和认识,例如,用一段文字对图像进行说明,在对文字或图像进行理解时,可以将文字中的关键词与图像中的相应区域对应起来,从而加深对图像或文字的理解[24]。为了达到上述目的,主要是要解决图像识别、图像语义标注、图像语义理解等问题,这 3 类技术也构成了面向图像型数据知识融合需要用到的主要技术,具体如高通滤波法[25]、小波变换融合增强法[26]等。

3.4 面向语音型数据的知识融合技术

在面向音频型数据的知识融合中,语音数据通常会与图形图像数据相结合,例如,在足球比赛的视频中,除了视频中人物动作和画面切换的识别之外,还可以针对进球时全场欢呼声进行分析,从而融合视频中声音的音量和音调等特征对进球片段进行知识提取。这一部分的技术需要使用语音特征提取、语音识别等技术。近年来,语音识别和转换的研究已经成绩斐然,在知识融合领域许多研究都是将语音转换为文字,利用文本知识融合的相关技术来实现对语音内容的融合。因此,计算层面向语音型数据的知识融合技术可分为语音处理与文本内容分析两个层次。语音处理技术部分除了包括常用的语音特征提取[27]、语音识别[28]等技术以外,还包括了相应的文本处理技术。

4 功能层知识融合技术

如前所述,知识融合的本质是满足人类对知识处理的功能需求或模拟人类知识处理的功能,具体地说就是通过计算实现知识形式的抽象和转换、发现知识

内容的相同与差异、解决知识化繁为简的问题以及找出知识中的联系与关系。因此可以将功能层中的知识融合技术主要分为面向知识转化的知识融合技术、面向知识细分的知识融合技术、面向知识比较的知识融合技术和面向知识依存的知识融合技术 4 个维度，如图 3 所示：

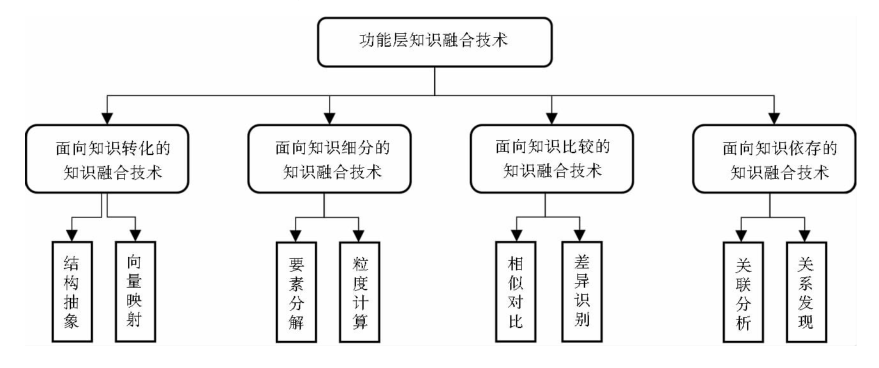


图 3　功能层知识融合技术体系

4.1　面向知识转化的知识融合技术

知识的转化是指在从原本的知识到融合后的知识这样一个过程中，知识必然会产生形态、结构和要素等方面的变化，将不同的知识通过抽象和映射转化到拥有统一尺度的空间之中，才能完成融合，这是知识融合的基础性工作。面向知识转化的知识融合技术可以进一步被细分成以下两种：

4.1.1　结构抽象

对知识结构的抽象就是将知识的不同要素之间的逻辑关系和位置关系表示出来，是对知识的物理结构及逻辑结构进行抽象和规则总结的过程，旨在便于将不同的知识放在同一个标准下进行知识融合的各项操作。知识融合中常用的结构抽象技术有完全句法分析、浅层句法分析以及依存句法分析等，具体如基于支持向量机的句法分析[29-30]、句法树功能库[31]。

4.1.2　向量映射

知识的向量映射就是将蕴含知识的原始数据中的特征用向量表示的方式转换为向量空间中的分量。借助知识的向量映射可以挖掘出知识特征及要素所蕴含的更深层次的语义信息，尤其是面向结构不明晰的知识时效果更好。知识向量映射对知识的融合效果与数据的规模和语义的丰富性成正比，即处理的数据量越大，知识涵义越多，最终融合得到的结果越有价值。知识融合中常用的向量映射技术法有向量空间模型、词袋模型以及概率语言模型等，具体如基于支持向量机的多特征融合来进行文本识别和分类[32]。

4.2　面向知识细分的知识融合技术

因为现实世界中的知识往往都具有整体性、系统性和多面性等特点，所以从细分角度来看知识融合的过程，可以将其理解为把不同来源知识中特定的要素提取出来并进行重新组合的过程，在这个过程中，就需要将原本的知识分解或粒化为更细致的单元知识。因此，对知识的细分主要是对知识进行分解得到知识要素或是对知识进行粒化得到知识颗粒。

4.2.1　要素分解

知识的要素分解是指在知识的物理和逻辑层面上以属性、特点和结构等为入口将知识分解成为一个个组成要素的形式，以便于人或机器选择相应的要素进行操作。知识要素分解主要可应用于知识融合中的自然语言处理、多类型知识的特征提取、知识图谱的应用、知识库的构建与补足等。常见的知识要素分解一般有基于词、基于短语、基于句子和基于篇章知识要素细分等。知识融合中常用的要素分解技术主要有数理统计、机器学习以及深度学习等，具体如基于过滤词典自动生成结合词汇密集度影响因子的术语抽取方法[33]。

4.2.2　粒度计算

知识的粒度计算在本质上是对知识中单元的分类，即通过分析知识的承载体中的数据结构来提取和发现隐藏在边界之中的知识颗粒。粒度计算在知识融合中可用于处理不确信的知识和多面性的知识，实现的过程和方法简单高效。常见的知识颗粒包括基于模糊规则下的知识颗粒和面向多层次粒度空间下的知识颗粒。知识融合中常用的粒度计算技术主要有基于粗糙集的粒计算、基于模糊集的粒计算、基于商空间的粒计算等，具体如颗粒度大小的分解与合并模型[34]。

4.3　面向知识比较的知识融合技术

知识比较是指通过一定的手段确定知识之间存在着的相同点和差异点。在知识融合中，知识要么具有相同的来源，要么是不同来源的同范畴知识，因

此知识间一定存在许多相同和差异的方面，知识比较就是要发现这些异同，包括知识的相似性和知识的差异性。

4.3.1 相似对比

知识的相似对比是指从知识特征的相似程度上来发现知识中的隐含规律以及知识内外部单元间的接近程度，以达到区分知识和发现知识的目的。知识的相似对比更多地强调的是语义的相似，而不仅仅是外在属性上的浅层相似，这是对知识进行深度挖掘和融合的必备条件。常见的知识相似度计算一般包括基于距离的相似对比、基于概率的相似对比和基于结构的相似对比等。常用的技术方法主要有距离计算、路径计算以及语义分析等，具体如特征词分布、LDA 主题分布、引文结构网络的相似计算[35]。

4.3.2 差异识别

知识的差异识别是指在对复杂的知识进行降维之后，通过对知识主要特征的比较，识别出知识间最为显著的差异点，以这些差异点作为知识区分或异常发现的标准。知识的差异识别尤其适用于有缺失值的知识或具有突变特征的知识，此外还适合结构性不强的知识或目的不太明确的知识融合任务。常见的知识差异识别包括离群知识的检测和突变知识的检测等。知识融合中常用的技术方法主要有主成分分析和因子分析等，具体如基于异常检测技术的技术机会识别[36]。

4.4 面向知识依存的知识融合技术

依存是指两个以上的事物同时存在并且相互依附和相互依赖。由于世界上的事物是普遍相关的，在知识融合的过程中，不同来源的知识通常会或多或少存在一定的关联，利用这些关联可以发现知识之间的关系，从而实现知识融合。

4.4.1 关联分析

知识的关联分析是指找寻存在于大量知识集合中的关联特性或相关特性，最终形成对知识中某些特质或属性同时出现的规律和规则的描述。在对知识进行处理之前，往往不知道其中可能蕴含的规律和规则，因此，知识关联分析更多的是一种数据驱动下的知识融合，能够发现有高价值的知识或是易被忽略的知识。常见的知识关联分析一般有基于规则的关联分析、基于概率的关联分析和基于路径的关联分析等。知识融合中常用的关联分析技术主要有关联规则挖掘、图模型以及概率统计模型等，具体如利用关联规则模型揭示共现关系背后的深层关联，帮助探测技术发展中潜藏的规律[37]。

4.4.2 关系发现

知识的关系发现是指以知识内部或外部的强相关关系为着手点挖掘出知识中存在的大量显性或隐性关系。知识关系发现的特点在于，在对知识进行处理之前已经有了大概的期望，提前预估了其中可能的关系或关系条件，因此，知识关系发现更多地属于业务驱动下的知识融合。常见的知识关系发现包括基于触发条件的关系发现和基于数理统计的关系发现。知识融合中常用的技术主要有逻辑语言、序列标注以及神经网络等，如非分类关系的抽取[38]。

5 任务层知识融合技术

任务层的知识融合技术在本质上是综合性的方法，它主要是各种计算层技术的组合，通过实现知识融合的各类功能，完成具体的知识融合任务。任务层的知识融合技术主要包括面向感知的知识融合技术、面向解释的知识融合技术、面向决策的知识融合技术和面向预测的知识融合技术，如图 4 所示：

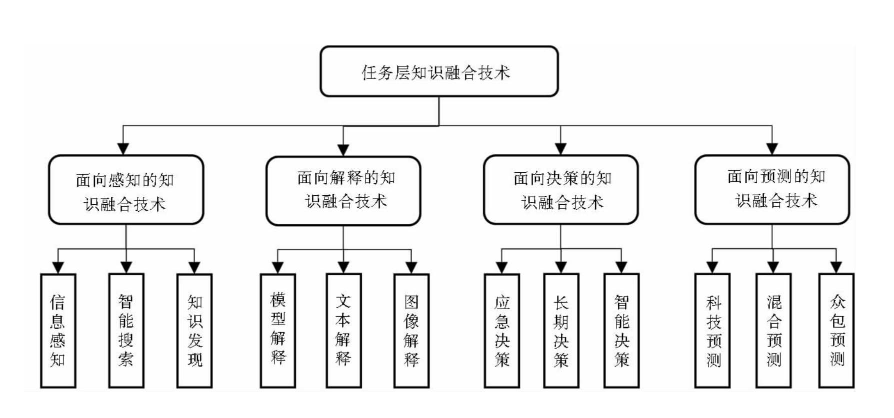


图 4 任务层知识融合技术体系

陈沫，李广建. 大数据环境下知识融合技术体系研究[J]. 图书情报工作,2022,66(20):20－31.

5.1 面向感知的知识融合技术

面向感知的知识融合技术，主要目的是解决数据和知识的发现、采集、获取，主要技术有信息感知、智能搜索和知识发现。

5.1.1 信息感知

信息感知通过综合获取同一对象的不同时空的数据或信息并加以融合，以多种信息网络的有效互联为基础，实现数据或信息的一体化和立体化采集与发现，从而得到相关的结论。数据和信息是知识融合所要处理的最基本对象，信息感知一直是知识融合领域的重要研究课题。知识融合中的信息感知技术通常基于线性代数、概率论、随机过程和统计计算、信息抽取、语义建模和自动聚类等，例如，在打击犯罪方面，已经可以综合使用上述技术从海量的 Web 信息源中感知并获取有价值的犯罪线索[39]。

5.1.2 智能搜索

智能搜索是搜索技术与知识融合技术相结合的产物，借助于情感分析、隐喻识别等知识融合技术在采集阶段就实现对信息内容的深度理解与信息场景的感知与识别，从而更加准确地过滤出与研究问题直接相关的有价值的信息或知识，减少所采集到的信息中的“噪音”，避免信息“过载”，甚至根据已经采集到的信息所涉及的场景，对信息采集的方向、范围等进行有意识的“优化”，主动发现表面上并不相关或者在开始采集规划阶段并没有考虑到的信息。

5.1.3 知识发现

知识发现是从各种信息中，根据不同的需求获得知识的过程。知识发现的目的是向使用者屏蔽原始数据的繁琐细节，从原始数据中提炼出有效的、新颖的、潜在有用的知识[40]。知识发现可以看作是信息感知和智能搜索的高级形式，知识发现所获得的知识是一种由数据、信息或知识到知识的高质量集成的结果。知识发现涉及 3 类技术：第一类是知识库建设技术[41]，第二类是知识的表示及其推理技术[42－44]，第三类是知识的应用技术[45]。

5.2 面向解释的知识融合技术

大数据环境下，来自多种领域和数据源的知识涉及不同学科，可能以不同的形式表达出来，因此缺乏一定的大众理解性，在知识融合的过程中，将不同学科知识通过相互解释从而达成理解是非常必要的。

5.2.1 面向数值解释的知识融合技术

数据融合是知识融合的重要组成部分，也是知识融合的一个基本阶段，数据融合大致可以分成两类，一类是非语义融合，另一类基于语义的融合。其中，非语义融合主要采用卡尔曼滤波器、贝叶斯网络和基于规则等技术[46]实现对数值形式的解释，而在基于语义的融合中，则多使用元数据技术、本体技术，将原始数据按照观察和测量的标准转换成相应的带有语义的描述格式，如 XML、RDF 等，以便帮助对数据含义的理解[47]。

5.2.2 面向文本解释的知识融合技术

这类技术是对大量分散的文本内容中的知识及其关系进行提取，根据已有的先验知识按一定的叙事逻辑从语义层面进行融合，从而帮助用户理解知识之间存在的因果关系。知识融合中文本语义解释主要涉及基于规则的、基于统计的以及基于机器学习的自然语言处理、信息提取、因果推断等技术，典型的应用有跨文档实体关系的抽取与合并[48－49]、语句融合排序[50]、文档主题聚类与摘要[51－52]等。

5.2.3 面向图像解释的知识融合技术

这类技术的目的是揭示出图像内容背后隐含的深层次语义，从而帮助用户对图像做出正确的理解。一般情况下，面向图像解释的知识融合在本质上是对图像进行语义融合，也就是将同一场景中具有互补信息和冗余特性的多个图像整合处理成为一个完整的、信息更丰富的、意义更为明确的图像，这种图像语义融合是一项综合性的工作，大致可以分成 3 个方面的内容：图像语义分割、图像语义标注以及基于语义的图像融合[53]。

5.3 面向决策的知识融合技术

面向决策的知识融合技术的目的是将来自多个来源的信息和知识整合成一些新的共同知识，这些知识可用于决策和问题解决，或可对所考虑的情况提供更好的洞察和理解。通常而言，可以从不同的角度对决策进行分类，这里根据知识融合技术在决策领域中的实际应用，重点讨论应急决策、长期决策以及智能决策任务中的知识融合技术。

5.3.1 面向应急决策的知识融合技术

应急决策就是决策者在有限时间内用已有知识对紧急事件的处置策略进行判断，并选择出满意的处理方案。在此类知识融合任务中，决策者利用已有的应对和处置突发事件知识，结合当前的情景信息，对需要决策的问题进行认识和理解，在此基础上形成决策。面向应急决策的知识融合通常会用到上下文识别[54]、模糊集合论[55]、情景分析[56]等技术。

5.3.2 面向长期决策的知识融合技术

长期决策是针对未来数十年或更长时间的潜在事

件进行的决策[57]。此类知识融合任务要对众多决策者的可能的评价标准进行提取、综合，从而选出可以用于确定决策解决方案的评价标准。目前这方面的典型做法是利用情感分析、模糊集合和熵值法等方法进行决策标准的融合[58-59]。

5.3.3 面向智能决策的知识融合技术

智能决策支持系统是人工智能和决策支持系统相结合的产物，是一种基于计算机的交互式辅助决策者进行决策的系统，其追求的目标是决策的自动化，减少决策对人的依赖，避免决策过程当中人的主观因素的干扰，统计分析技术、知识抽取、知识表示、知识比较、关系发现等技术在这个过程中发挥着重要的作用，并且有成功的应用[60]。

5.4 面向预测的知识融合技术

大数据环境下，知识融合开始直接参与到问题解决和复杂决策之中。在理论上，相关学者已经证明了基于多源信息和知识融合以及群体智慧的、由信息驱动的预测，是一种非常有效的预测方式[61-62]。因此，近年来，学界和业界都十分重视采用知识融合的思想解决预测问题。

5.4.1 面向技术预测的知识融合技术

面向技术预测的知识融合技术是指利用知识融合来实现对技术发展的预测。例如，K. Heyeol 提出基于多源数据融合的新兴科技情报分析框架[63]，该框架综合利用传统科学出版物以及 Web 学术数据、Wikipedia 数据等，通过潜在语义分析、知识提取，实现了对新兴技术的分析和预测。

5.4.2 面向人机混合预测的知识融合技术

面向人机混合预测的知识融合技术是指在预测过程中使机器概率化的智能与预测，分析人员有机地参与到智能交互的知识融合过程中，与机器行为互为补充，共同协作完成复杂的预测工作。这方面，DARPA 和南加州大学的研究人员都做过有益的尝试，取得了很好的效果[64-65]。

5.4.3 面向众包预测的知识融合技术

面向众包预测的知识融合技术是指在预测任务中利用群体智慧效应获得相对最优的预测结果。其本质就是将分析者的智慧、知识、信息通过网络平台进行聚合及融合并通过市场行为形成最优化的预测结果，这方面的典型代表是 P. E. Tetlock 等[66]提出的聚合群体智慧进行预测的技术方法体系，该技术方法体系强调从多种来源收集证据、注重概率思维、强调团队合作、多因素分析以及持续改进等技术方法。

6 结语

大数据时代信息技术的发展及其应用，极大丰富了知识融合的技术方法与体系，本文从大数据时代知识融合的多源、异构的数据特点出发，结合大数据技术的生态，分析了大数据环境下知识融合技术的内涵，对知识融合的特点及其实现的关键技术进行了深入分析，以知识融合实现过程中的功能为中介，综合考虑各种技术的自身特性、适用的知识对象、应用的抽象层次，建立了具有计算层、功能层和任务层 3 个层次的技术体系架构，并对每一层次的知识融合技术进行概括和总结。其中，计算层的知识融合技术是最为具体的技术，反映了各种技术的可计算性、可度量性和可检验性，表现为算法流程、实施模型、实现公式等。功能层的知识融合技术着眼于它们在知识融合过程中发挥的作用，任何一项知识融合工作，都需要借助具体的技术，实现一定的功能，这种功能实际上是各种以数据（知识）为处理对象的具体技术所能发挥作用的形式化的描述，计算层的技术在本质上是“底层”的可操作的技术，严格地说，它只能处理“数据”，不能处理知识，而功能层的技术，则是在计算层技术数据处理的技术上，将“数据”处理上升为“知识”处理的技术，为解决知识融合任务奠定基础。

本文在前人研究的基础上，将知识融合的功能总结为知识转化、知识细分、知识比较和知识依存，并进一步根据它们的特质将所涉及的主要技术领域划分为结构抽象与向量映射、要素分解与粒度计算、相似对比与差异识别、关联分析与关系发现等方面，论述了各个方面中的代表性技术。知识融合功能的划分及其在整个知识融合技术体系中的引入，是文本知识融合技术体系与以往知识融合技术体系研究的显著不同点，功能层在整个知识融合技术体系中起着承上启下的作用，解释了知识融合过程中处理对象从“数据”到“知识”的转化过程。任务层的知识融合技术，反映了解决知识融合问题的逻辑思维过程，表现为知识融合计算层的各种技术和/或功能层的各种技术的组合，其本质上是为了实现知识融合的目标，解决知识融合的问题，具有明显的针对性和明确的目的性。在任务层中，本文将知识融合的任务明确地抽象为感知、解释、决策和预测，这也是本文的一个特色。

本文构建的知识融合技术体系，是一种多层、多维立体的知识融合技术，弥补了以往知识融合技术体系研究中多就技术原理这一特性进行分类，较少考虑技

术应用场景的不足,特别是本文构建的三层知识融合技术体系,各层次相互联系、相互影响、环环相扣,向上可以抽象,可以与知识融合的具体问题和具体任务相关联,向下可以具化,可以指引找到进行知识抽取、采集、组织、建模、组合、推理、集成的可操作、可计算的技术方法,它除了对知识融合过程中涉及的技术具有认识论意义,还可以对知识融合技术的选型提供借鉴,指导知识融合的实践工作。

参考文献:

[1] 董小英. 知识优势的理论基础与战略选择[J]. 北京大学学报(哲学社会科学版),2004(4):37－45.
[2] YU X, LIN Q. Knowledge fusion methods: a survey[C]// Proceedings of the 2017 2nd international conference on software, multimedia and communication engineering. Colorado: DEStech Publications, 2017: 300－304.
[3] 邱均平,余厚强. 知识科学视角下国际知识融合研究进展与趋势[J]. 图书情报工作,2015,59(8):126－132,148.
[4] 缐锦. 知识融合中若干关键技术研究[D]. 杭州:浙江大学,2005.
[5] 姜振寰,吴明泰,王海山,等. 技术学辞典[M]. 沈阳:辽宁科学技术出版社,1990.
[6] 杨开城,王斌. 从技术的本质看教育技术的本质[J]. 中国电化教育,2007(9):1－4.
[7] 徐敏. 知识计算视角下知识融合技术的模式及方法研究[D]. 北京:北京大学,2019.
[8] DONG X L, GABRILOVICH E, HEITZ G, et al. From data fusion to knowledge fusion[J]. arXiv preprint arXiv: 1503. 00302, 2015. [2022－09－22]. https://arxiv. org/pdf/1503. 00302v1. pdf.
[9] ROEMERr M J, KACPRZYNSKI G J, ORSAGH R F. Assessment of data and knowledge fusion strategies for prognostics and health management[C]//2001 IEEE aerospace conference proceedings. New York: IEEE, 2001: 2979－2988.
[10] PREECE A, HUI K, GRAY A, et al. KRAFT: an agent architecture for knowledge fusion[J]. International journal of cooperative information systems, 2001, 10(1/2): 171－195.
[11] TSUKASA I, TAKENAKA T, MOTOMURA Y. Customer behavior prediction system by large scale data fusion in a retail service[J]. Transactions of the Japanese Society for Artificial Intelligence, 2011, 26(6): 670－681.
[12] FENG G, ZHANG J D, LIAO S S. A novel method for combining Bayesian networks, theoretical analysis, and its applications[J]. Pattern recognition, 2014, 47(5): 2057－2069.
[13] 周芳,韩立岩. 基于知识融合的公司失败判别方法[J]. 财会通讯,2015(8):61－63.
[14] SAHA R K, CHANG K C. An efficient algorithm for multisensor track fusion[J]. IEEE transactions on aerospace and electronic systems, 1998, 34(1):200－210.
[15] FENG G, ZHANG J D, LIAO S S. A novel method for combining Bayesian networks, theoretical analysis, and its applications[J]. Pattern recognition, 2014, 47(5): 2057－2069.
[16] MARTENS D, DEBAKER M, HAESEN R, et al. Ant-based approach to the knowledge fusion problem [C]// International workshop on ant colony optimization and swarm intelligence. Berlin: Springer, 2006: 84－95.
[17] 马冯. 基于扩展概念格的多数据源分类知识融合问题研究[D]. 合肥:合肥工业大学,2006.
[18] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion[C]//Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2014: 601－610.
[19] TAI C H, CHANG C T, CHANG Y S. Hybrid knowledge fusion and inference on cloud environment [J]. Future generation computer systems, 2018, 87(1): 568－579.
[20] GRUBER T R. Ontolingua: a mechanism to support portable ontologies [R]. California: Knowledge Systems Laboratory, 1992.
[21] GOU J, JIANG Y, WU Y, et al. A New knowledge fusion method based on semantic rules[C]// International conference on signal processing. New York: IEEE, 2006: 1939－1942.
[22] 张灵凯,于良. 多源遥感数据融合研究综述[J]. 城市地理,2017(2):173.
[23] 张孝飞,王强,韦春荣,等. 医学图像融合技术研究综述[J]. 广西科学,2002(1):64－68.
[24] KARPATHY A, LI F F. Deep visual-semantic alignments for generating image descriptions[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 9(4): 664－676.
[25] CHAVEZ P S, SILDES S C, ANDERSON J A. Comparison of three different methods to merge multiresolution and multispectral data: landsat TM and SPOT panchromatic[J]. Photogrammetric engineering & remote sensing, 1991, 57(3): 295－303.
[26] LI H, MANJUNATH B S, MITRA S K. Multisensor image fusion using the wavelet transform. graph[J]. Models image process, 1995, 57(3): 235－245.
[27] 李琦,孙桂玲,黄翠,等. 基于水声环境空间中多模态深度融合模型的目标识别方法研究[J]. 海洋技术学报. 2019,38(6):35－45.
[28] 胡郁. 人工智能与语音识别技术[J]. 电子产品世界,2016,23(4):23－25,27.
[29] BRAVO A, CASES M, QUERALT-ROSINACH N, et al. A knowledge-driven approach to extract disease-related biomarkers from the literature[R]. BioMed research international, 2014: 1－11.
[30] BRAVO À, PINERO J, QUERALT-ROSINACHh N. et al. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research[J]. BMC bioinformatics, 2015, 16(1): 1－17.

[31] 陈建美,林鸿飞. 中文情感常识知识库的构建[J]. 情报学报, 2009,28(4):492-498.
[32] 龚安,费凡. 基于多特征融合的评论文本情感分析[J]. 计算机技术与发展,2018(8):91-95.
[33] 曾镇,吕学强,李卓. 一种面向专利摘要的领域术语抽取方法[J]. 计算机应用与软件,2016,33(3):48-51.
[34] HOBBS J R. Granularity[C] //Proceedings of the ninth intenational joint conference on artificial intelligence. Los Angeles: Morgan Kaufmann, 1985: 432-435.
[35] 丁梦晓,毕强,许鹏程,等. 基于用户兴趣度量的知识发现服务精准推荐[J]. 图书情报工作, 2019,63(3):21-29.
[36] 翟东升,郭程,张杰,等. 采用异常检测的技术机会识别方法研究[J]. 现代图书情报技术,2016(10):81-90.
[37] 周磊,杨威. 基于加权关联规则的技术融合探测[J]. 情报杂志, 2019,38(1):67-72.
[38] 朱惠,王昊,苏新宁,等. 汉语领域术语非分类关系抽取方法研究[J]. 情报学报,2018,37(12):1193-1203.
[39] KEJRIWAL M, SZEKELY P, KNOBLOCK C. Investigative knowledge discovery for combating illicit activities[J]. IEEE intelligent systems, 2018 (1): 53-63.
[40] 陆雄文. 管理学大辞典[M]. 上海:上海辞书出版社,2013.
[41] LENAT D B, PRAKASH M, SHEPTHRD M. CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks[J]. AI magazine, 1985, 6(4): 65-65.
[42] 刘建炜,燕路峰. 知识表示方法比较[J]. 计算机系统应用, 2011,20(3):242-246.
[43] 黄德根,张云霞,林红梅,等. 基于规则推理网络的分类模型[J]. 软件学报,2020,31(4):1063-1078.
[44] CHEN Y W, YANG J B, XU D L, et al. Inference analysis and adaptive training for belief rule based systems[J]. Expert systems with applications, 2011, 38(10): 12845-12860.
[45] YAN R, LI G, LIU B. Knowledge fusion based on DS theory and its application on expert system for software fault diagnosis[C]// 2015 prognostics and system health management conference. New York: IEEE, 2015: 1-5.
[46] PARTESCANO E, BROSICH A, LIPIZER M, et al. From heterogeneous marine sensors to sensor Web: (near) real-time open data access adopting OGC sensor Web enablement standards[J]. Open geospatial data, software and standards, 2017, 2(1): 1-9.
[47] SHETH A, HENSON C, SAHOO S S. Semantic sensor Web[J]. IEEE Internet computing, 2008, 12(4): 78-83.
[48] LI X, MA S, ZHOU X. Large-scale Chinese cross-document entity disambiguation and information fusion[C]// Advancing big data benchmarks. Berlin: Springer, 2013: 105-119.
[49] BRONSELAER A, VAN Britsom D, DE T G. A framework for multiset merging[J]. Fuzzy sets and systems, 2012, 191(1): 1-20.
[50] YUE L, SHI Z, HAN J, et al. Multi-factors based sentence ordering for cross-document fusion from multimodal content[J]. Neurocomputing, 2017, 253(1): 6-14.
[51] WITTE R, BERGLER S. Fuzzy clustering for topic analysis and summarization of document collections[C]//Conference of the canadian society for computational studies of intelligence. Berlin: Springer, 2007: 476-488.
[52] LEBANOFF L, WANG B, FENG Z, et al. Modeling endorsement for multi-document abstractive summarization[J]. arXiv preprint. arXiv: 2110. 07844, 2021.
[53] 倪景秀. 图像语义融合关键技术的研究[D]. 北京:中国矿业大学,2018.
[54] SMIRNOV A V, LEVASHOVA T, SHILOV N. Knowledge fusion in context-aware decision support systems[C]// Proceedings of the international joint conference on knowledge discovery, knowledge engineering and knowledge management. Rome: KEOD, 2014: 186-194.
[55] 张磊. 具有模糊不确定性的应急决策知识融合方法研究[D]. 大连:大连理工大学,2019.
[56] 张志霞, 郝纹慧. 基于知识元的突发灾害事故动态情景模型[J]. 油气储运, 2019, 38(9): 980-987, 995.
[57] The RAND Corporation. Seeking examples of long-term decisions [EB/OL]. [2022-08-20]. https://www.rand.org/pardee/LongTermDecisions/seeking.html.
[58] ÇALI S, BALAMAN Ş Y. Improved decisions for marketing, supply and purchasing: mining big data through an integration of sentiment analysis and intuitionistic fuzzy multi criteria assessment[J]. Computers & industrial engineering, 2019, 129(1): 315-332.
[59] CHEN X, ZHANG W, XU X, et al. A public and large-scale expert information fusion method and its application: mining public opinion via sentiment analysis and measuring public dynamic reliability[J]. Information fusion, 2022, 78(1): 71-85.
[60] MORGE M. The hedgehog and the fox[C]//International workshop on argumentation in multi-agent systems. Berlin: Springer, 2007: 114-131.
[61] CHARBONNEAU S, FYE S, HAY J, et al. A retrospective analysis of technology forecasting [C/OL] //AIAA SPACE 2013 conference and exposition. California: American Institute of Aeronautics and Astronautics, 2013[2022-06-14]. https://doi.org/10.2514/6.2013-5519. DOI:10.2514/6.2013-5519.
[62] PAGE S E. The difference: how the power of diversity creates better groups, firms, schools, and societies-new edition[M]. New Jersey: Princeton University Press, 2008[2022-06-14]. https://doi.org/10.2514/6.2013-5529. DOI: 10.2514/6.2013-5519.
[63] HEYEOL K. Data-driven technology foresight: text analysis of emerging technologies[D]. Seoul: Seoul National University, 2018.
[64] DARPA and data: a portfolio overview[EB/OL]. [2022-08-13]. https://www.nitrd.gov/nitrdgroups/images/3/31/DARPA-and-DATA.pdf.
[65] MORSTATTER F, GALSTYAN A, SATYUKOV G, et al. SAGE:

a hybrid geopolitical event forecasting system[C]//Proceedings of the 28th international joint conference on artificial intelligence. California: AAAI Press, 2019: 6557－6559.
[66] TETLOCK P E, GARDNER D. Super forecasting: the art and science of prediction[M]. New York: Random House, 2016.

作者贡献说明：

陈沫：相关文献的收集、整理与分析，撰写和修改论文；
李广建：提出研究问题和大纲，撰写和修改论文。

Research on the Knowledge Fusion Technology Taxonomy in Big Data Environment

Chen Mo　Li Guangjian
Department of Information Management, Peking University, Beijing 100871

Abstract: [**Purpose/Significance**] This paper mainly studies the key technologies of multi-source knowledge fusion in the big data environment, and proposes a complete set of technology taxonomy based on the characteristics of multi-source knowledge objects in different fields to provide technical support and solutions for the realization of knowledge fusion. [**Method/Process**] The study utilized qualitative analysis method to analyze the existing related research, and then within the same hierarchical level on the related contents of induction and deduction, used literature analysis method, to solve the problem of combing knowledge fusion, sum up the knowledge integration of task type and implement various tasks involved in the work process and its specific technology, and form a knowledge fusion system. [**Result/Conclusion**] Considering all kinds of technology's own characteristics, applicable knowledge objects, the application of abstraction of knowledge, the paper establishes the calculating layer, function layer and mission layer - three levels of technical architecture. These three layers contact each other, influence each other and interlock each other. The upper layer can be abstracted and associated with specific problems (tasks) of knowledge fusion. The lower layer can be embodied, that is, to find operational and computable technical methods to solve specific problems of knowledge fusion.

Keywords: knowledge fusion　technology system　big data



2021 图书馆馆长论坛